ELSEVIER

Theoretical Computer Science 292 (2003) 345–357

Theoretical Computer Science

www.elsevier.com/locate/tcs

Formal logics of discovery and hypothesis formation by machine

Petr Hájek*, Martin Holeňa

Institute of Computer Science, Academy of Sciences, Pod vodárenskou věží 2, CZ-182 07 Prague, Czech Republic

Abstract

The following are the aims of the paper: (1) To call the attention of the community of Discovery Science (DS) to certain existing formal systems for DS developed in Prague in the 1960s through the 1980s suitable for DS and unfortunately largely unknown. (2) To illustrate the use of the calculi in question by the example of the GUHA method of hypothesis generation by computer, subjecting this method to a critical evaluation in the context of contemporary data mining. (3) To stress the importance of fuzzy logic for DS and to present the state of mathematical foundations of fuzzy logic. (4) Finally, to present a running research program of developing calculi of symbolic fuzzy logic for DS and for a fuzzy GUHA method.

Keywords: Logic of discovery; Hypothesis generation; GUHA method; Data analysis; Data Mining; Fuzzy logic

1. Introduction

The term "logic of discovery" is admittedly not new: let us mention at least Popper's philosophical work [50], Buchanan's dissertation [3] analysing the notion of a logic of discovery in relation to artificial intelligence and Plotkin's paper [49] with his notion of a logic of discovery as a logic of induction plus a logic of suggestion. In relation to data mining one has to mention the concept of exploratory data analysis, as elaborated by Tukey [58]. Can there be a formal (symbolic) logic of discovery? And why should it be developed? The answer is yes, various formal calculi can be and have been developed.

* Corresponding author.
E-mail addresses: hajek@cs.cas.cz (P. Hájek), martin@cs.cas.cz (M. Holeňa).
URLs: http://www.cs.cas.cz/∼hajek/, http://www.cs.cas.cz/∼martin/

0304-3975/03/$ - see front matter
PII: S0304-3975(02)00175-5

And the obvious *raison d'être* for them (besides their purely logical importance) is that the computer can understand, process and (sometimes) evaluate formulas of a formal language, which is important for discovery as a cognitive activity studied by AI and Discovery Science (DS). The present paper has the following aims: (1) To call the attention of the DS community to certain existing formal systems for DS developed in Prague in the 1960s through the 1980s—not just for some reasons of priority but since we find them natural, suitable for DS and unfortunately largely unknown. (2) To illustrate the use of the calculi in question by the example of the GUHA method of hypothesis generation by computer, subjecting this method to a critical evaluation in the context of contemporary data mining. (3) To stress the importance of fuzzy logic (and, more generally soft computing) for DS and to present the state of mathematical foundations of fuzzy logic. (4) Finally, to present a running research program of developing calculi of symbolic fuzzy logic for DS and for a fuzzy GUHA method.

2. Calculi of the logic of discovery

We refer here on calculi whose syntax and semantics is fully elaborated in the monograph [19]. Since there exists a survey paper [18] (which we would like to recommend to the reader) we shall be rather sketchy (see also [12]).

Distinction is made between observational and theoretical languages. Formulas of an observational language are used to speak about the data; formulas of a theoretical language on a universe not directly being at our disposal. For example, it may happen that out of 100 randomly taken samples of river ground, 50 were sandy (observational). However, still actually only 30% of the ground of that river may be sandy (theoretical).

Data are finite structures. For simplicity, think of a data structure as of a rectangular matrix whose rows correspond to objects and columns correspond to values of a variate. Each variate has a *name* (X, TEST, etc.) and domain from which the values are taken (real, integer, numbers 1–20, etc.). Names of variates are called *predicates*. We may also have some distinguished subsets of the domain capturing particular concepts, e.g. the interval $\langle 10, 20\rangle$. The fact that the value of a variate lies in some particular distinguished subset can be expressed using an *atomic formula*, such as

`glowable proportion < 10%`,
`number of trees within 5 m distance = 2`,
`Tubificidae = yes`, or more simply, `Tubificidae`.

From atomic formulae, general *open formulae* can be built by means of the usual logical connectives ($\neg$, $\wedge$, ...). In particular, *conjunctive formulae* express some combined property (combination of single properties), of an individual object, e.g.,

`glowable proportion < 10%` $\wedge$ `grain diameter > 1.5 mm` $\wedge \wedge$ `water level < 0.4 m`.

Given data $\mathbf{M}$, it is clear what we mean saying that an object m *satisfies* an open formula φ; $Fr_{\mathbf{M}}(\varphi)$ is the frequency of φ in $\mathbf{M}$, i.e. the number of objects in $\mathbf{M}$ satisfying φ. For a pair φ, ψ of open formulas we have four frequencies $a_{\mathbf{M}} = Fr_{\mathbf{M}}(\varphi \& \psi)$, $b_{\mathbf{M}} = Fr_{\mathbf{M}}(\varphi \& \neg\psi)$, $c_{\mathbf{M}} = Fr_{M}(\neg\varphi \& \psi)$, $d_{\mathbf{M}} = Fr_{\mathbf{M}}(\neg\varphi \& \neg\psi)$ ($\neg$ being negation, & being conjunction). Finally, the quadruple $(a_{\mathbf{M}}, b_{\mathbf{M}}, c_{\mathbf{M}}, d_{\mathbf{M}})$ is called the *four-fold table* of (φ, ψ) in M.

Generalized quantifiers are used to form sentences, i.e. formulas expressing properties of the data as whole. For example, the fact that we deal with data about adult patients can be expressed as $(\forall x)$`age > 15`. Similarly, the fact that in many (say, at least 90%) of the considered samples of fauna the family Oligochaeta was present can be expressed as $(\mathrm{Many}_{90\%}x)$ `Oligochaeta`, the formula Many_p meaning $Fr(\varphi)\geqslant p$. In the case of binary quantifiers or other quantifiers of a higher arity, the closed formula $(Qx)(\varphi_1,\ldots,\varphi_m)$, built from an m-ary generalized quantifier Q and open formulae $\varphi_1,\ldots,\varphi_m$, in general states some relationship between properties corresponding to $\varphi_1,\ldots,\varphi_m$. As an example can serve the quantifier $(\mathrm{Many}_p x)\ (\psi\mid\varphi)$, written also $\varphi\sqsupset_p\psi$, which means $Fr(\varphi\,\&\,\psi)/Fr(\varphi)\geqslant p$. Other examples are the quantifiers $\sqsupset^{!}_{\alpha,\theta}$ and $\sqsupset^{?}_{\alpha,\theta}$, introduced below.

The semantics of a unary quantifier q is given by its *truth function* (also called *associated function*) Tr_q assigning to each column vector of zeros and ones (the course of values of a formula) 0 or 1. For example, $Tr_{\mathrm{Majority}}(V)=1$ iff the column V contains more 1's than 0's. Similarly for a binary quantifier (like "Many ... are ..."), but now V is a matrix consisting of two column vectors of 0's and 1's.

One of most distinguishing features of the described approach is a *tight connection between logic and statistic*, a connection whose importance has been rediscovered two decades later in the context of modern data mining [38,39,62,64]. The key idea of that connection is to view each data matrix, used for evaluating observational sentences, as a realization of a random sample. Consequently, the truth function of a generalized quantifier, composed with random samples with values in its domain, is a random variable. Since random variables expressible as a composition of a function of many variables with multidimensional random samples are often used as test statistics for *testing statistical hypotheses*, it is possible to cast statistical tests in the framework of generalized quantifiers. In the most simple case of dichotomous data matrices, this can be accomplished for example as follows: Let M_D be a two-column matrix of zeros and ones the rows of which contain evaluations, in given data, of some pair of open formulae (φ,ψ). Thus all those evaluations are viewed as realizations of independent two-dimensional random vectors, all having the same distribution D. Suppose that D is known to belong to the set $\mathscr{D}$ described by the nonsingularity condition $p_{\psi|\varphi}\in(0,1)$, where $p_{\psi|\varphi}$ is the conditional probability corresponding to D of ψ being satisfied conditioned on φ being satisfied. The parametrizability of $\mathscr{D}$ by $p_{\psi|\varphi}$ makes it possible to express also a null hypothesis $D\in\mathscr{D}_0$ by means of $p_{\psi|\varphi}$. In particular, given $\alpha,\theta\in(0,1)$, the following statistical test can be considered: *test the null hypothesis* $p_{\psi|\varphi}\leqslant\theta$ *using a test statistic* $\sum_{i=a}^{a+b}\binom{a+b}{i}\theta^i(1-\theta)^{a+b-i}$, *and the critical region* $(0,\alpha\rangle$. This leads to a binary quantifier $\sqsupset^{!}_{\alpha,\theta}$ called *lower critical implication* (lci) with the threshold θ, whose truth function is defined, for each natural k and each matrix $M\in\{0,1\}^{k,2}$, as follows:

$$\mathrm{Tr}_{\sqsupset^{!}_{\theta}}(M)=1 \quad \text{iff} \quad \sum_{i=a}^{a+b}\binom{a+b}{i}\theta^i(1-\theta)^{a+b-i}\leqslant\alpha.$$

Thus the quantifier $\sqsupset^{i}_{\alpha,\theta}$ captures the fact that the test leads to rejecting $p_{\psi|\varphi}\leqslant\theta$ at the significance level α.

Dually, we may consider a quantifier (upper critical implication) capturing the fact that a particular test does not reject $p_{\psi|\varphi} \geqslant \theta$. It is important that both quantifiers belong to an infinite family of implicational (multitudinal) quantifiers defined by simple monotonicity conditions (a formal definition is given below in Section 3). All quantifiers $\sqsupset$ from this family share some important logical properties (e.g., $(\varphi_1 \,\&\, \varphi_2) \sqsupset \psi$ implies $\varphi_1 \sqsupset (\neg\varphi_2 \vee \psi)$).

In this way we get observational logical calculi with interesting formal properties—a particular branch of *finite model theory* as logical foundations of database theory. Sentences of an observational language express interesting *patterns* that may be recognized in given data. In contrast, *theoretical sentences* are interpreted in possibly infinite structures, not directly accessible. They may express properties of *probability, possibility*, (*in*) *dependence*, *etc*. The corresponding calculi have been elaborated, also using the notion of a generalized quantifier. *Modal logic* is relevant here as theoretical structures are defined as parametrized by "possible worlds" and e.g. the probability of φ is defined as the probability of the set of all possible worlds in which φ is true. See the references above for details and note that there is important literature on probability quantifiers, notably [37] and, in our context, [13]. An early paper on complexity problems of calculi of the described kind is [51].

Inductive inference is the step from an observation (expressed by a sentence α of an observational language) to a theoretical sentence Φ, given some theoretical frame assumption Frame. The rationality of such step is given by the fact that assuming Frame, if Φ were false then we could prove that the observation α is unlikely (in some specified sense), i.e.

$$\text{Frame}, \neg\Phi \vdash \text{unlikely}(\alpha).$$

This is a starting point for various formal developments, including (but not identical with) statistical hypothesis testing.

3. The GUHA method and data mining

The development of this method of exploratory data analysis started in mid-1960s by papers by Hájek et al [17]. Even if the original formalism appears simple-minded today, the principle formulated there remains valid, *namely*: use means of formal logic to let the computer generate all hypotheses interesting with respect to a research task and supported by the data. In fact the computer generates interesting observational sentences rather than hypotheses (theoretical sentences); but the observational sentences correspond to theoretical sentences via a rule of inductive inference as above and, in addition, they are interesting as statements about the data themselves, in particular if the data are immensely large.

This general program may be realized in various forms. The main form that has been implemented (repeatedly) and practically used is the GUHA package ASSOC for generating hypotheses on associations and high conditional probabilities using various binary quantifiers of two particular kinds—associational and implicational [21,23,53,54].

- A binary generalized quantifier is called *associational* if the value of its truth function $Tr_\sim$ is fully determined by the four-fold table (a_M, b_M, c_M, d_M) of the considered formulae in a particular data M (i.e., $Tr_\sim : \mathcal{N}_0^4 \to \{0,1\}$) and the following holds for each pair M, M':

$$\text{IF } a_{M'} \geqslant a_M \ \&\ b_{M'} \leqslant b_M \ \&\ c_{M'} \leqslant c_M \ \&\ d_{M'} \geqslant d_M \ \&\ Tr_\sim(a_M, b_M, c_M, d_M) = 1, \text{ THEN } Tr_\sim(a_{M'}, b_{M'}, c_{M'}, d_{M'}) = 1. \quad (1)$$

  From (1) follows that the associational quantifier $\sim$ in a sentence $\varphi \sim \psi$ is suitable to capture statistical tests based on a high empirical correlation between the occurrence in the data of the properties expressed by the formulae φ and ψ, such as tests of independence. Examples of associational quantifiers are the *Fisher quantifier* $\sim_\alpha^F$, corresponding to the one-sided Fisher exact test of independence in four-fold tables with the significance level $\alpha \in (0,1)$, and the *chi-square quantifier* $\sim_\alpha^{\chi^2}$, corresponding to the χ^2 asymptotic test of independence in four-fold tables with the significance level α.
- A binary generalized quantifier $\sqsupset$ is called *implicational* (*multitudinal*) if the value of its true function $Tr_\sqsupset$ is fully determined by the four-fold table (a_M, b_M, c_M, d_M) and the following holds for each pair M, M':

$$\text{IF } a_{M'} \geqslant a_M \ \&\ b_{M'} \leqslant b_M \ \&\ Tr_\sqsupset(a_M, b_M, c_M, d_M) = 1, \text{ THEN } Tr_\sqsupset(a_{M'}, b_{M'}, c_{M'}, d_{M'}) = 1. \quad (2)$$

  Due to (2), the implicational quantifier $\sqsupset$ in a sentence $\varphi \sqsupset \psi$ is suitable to capture tests based on a high relative frequence of objects with the property expressed by ψ among objects with the property expressed by φ, such as the binomial test. Examples of implicational quantifiers are the generalized quantifiers $\sqsupset_p$, $\sqsupset_{\alpha,\theta}^i$, $\sqsupset_{\alpha,\theta}^?$ from Section 2.

Observe that condition (2) is stronger than (1) since it assures transferring the truth from the table (a_M, b_M, c_M, d_M) to the table (a'_M, b'_M, c'_M, d'_M) without making any assumptions about c'_M and d'_M. Consequently, each implicational quantifier is also associational. For a theoretical analysis of those kinds of quantifiers, the reader is referred to the monograph [19], for recent developments see [23,54].

At the time when the theoretical principles of the GUHA approach were developed, data analysts typically dealt with tens to hundreds of objects, with thousands being already exceptional. Future databases of a size breaking the terabyte limit will increasingly often contain data about the whole population, thus making the inference from a sample to the population in principle superfluous. Indeed, sample-based methods are used in data mining mostly for efficiency reasons [42,44,45,57,61]. But notice that an inductive inference from frame assumptions and an observation to a theoretical sentence is not an inference from a sample to the population. Therefore, it remains fully justified even if the observation is based on data covering the whole population, i.e. even under the conditions of data mining.

GUHA has been developed as a method of exploratory data analysis but has been rather rarely used in practice until recently, when the availability of a PC version of ASSOC and the ubiquity of personal computers has made its sophisticated algorithms easily accessible even for occasional users needing to tackle realistically-sized problems. Indeed, since 1995 more than a dozen GUHA applications have been reported in various application domains, such as medicine [46,63], pharmacy [22] (in Czech–Japanese cooperation), economy [47,48], or musicology [7].

We would like to illustrate the typical use of GUHA with a recent application in the area of *river ecology* [30,29]. One of the very efficient ways to increase the suitability of rivers for water transport is building *groynes*. On the other hand, ecologists often fear the changes in the biocoenosis of the river and its banks to which groynes may lead. However, it is a matter of fact that the complex relationships between the biocoenosis and the ecological factors characterizing a groyne field are only poorly understood so far. Therefore, a research project has been launched in Germany 1998, with the objective to investigate those relationships, and to propose an empirically proven model capturing them and allowing to estimate the changes in the biocoenosis that prospective groynes would cause. That model and the knowledge on which it will be based are intended to serve as a basis for constructing, at the end of the project, a decision support system for river ecology.

The project has been founded by the German federal ministry of education and research, and is being accomplished by researchers from four institutions—the German universities of Cottbus, Darmstadt and Marburg, and the Czech Academy of Sciences. On the important Czech and German river Elbe, five groyne fields have been chosen, and a large amount of empirical data on them has been collected during 1998–1999. The main part of the collected data is formed by nearly 1000 field samples of *aquatic fauna* and more than 1400 field samples of *terrestrial fauna*. Each sample includes all animals caught in special traps during some prescribed period of time, ranging from several hours to 2 days. Simultaneously with collecting those samples, various *ecological factors* have been measured in the groyne fields, e.g., oxygen concentration, diameter of ground grains, glowable proportion of the ground material, whereas others, such as water level and flow velocity, have been computed using a hydrodynamic simulation model.

The collected data are, first of all, analysed with respect to the species contained in them. Then some preprocessing is performed, and finally data mining methods, including GUHA, are applied to the preprocessed data. Fig. 1 shows an example of results obtained when simultaneously applying the Fisher quantifier and the quantifier lci to those data. So far, only a small fraction of the collected data has been analysed with respect to the contained species. Moreover, the order in which the samples of fauna are being analysed with respect to the contained species is determined mainly by the order in which they were collected, and is not independent of the values of the considered ecological factors. Therefore, the first obtained results have only a limited value from the point of view of the objectives of the whole research project. Nevertheless, they already confirm that data mining can yield interesting knowledge about the relationships between ecological factors, and aquatic and terrestrial fauna in groyne fields.

DB "carab" analysis: non-dichotomously present species + 1-5 ecological factors

shortest hypotheses with support 4 by Fisher quantifier and lci with p = 1/2, significance level 10 %

Ecological factors					Species	Fisher quantifier	lower critical implication
sand = 1	height herbs = > 40				Pseudoophonus_rufipes = 1 - 2	1.2e-005	0.073
distance WL = > 70	distance FG = < 30	cover litter = < 20			Formicidae = yes	1.2e-005	0.063
distance WL = > 70	distance FG = < 30	cover litter = < 40			Formicidae = yes	1.2e-005	0.063
distance WL = > 70	distance FG = < 30	cover litter = < 50			Formicidae = yes	1.2e-005	0.063
distance WL = > 70	distance FG = < 30	cover litter = < 60			Formicidae = yes	1.2e-005	0.063
distance WL = > 70	distance FG = -50 - 30	cover litter = < 20			Formicidae = yes	1.2e-005	0.063
distance WL = > 70	distance FG = -50 - 30	cover litter = < 40			Formicidae = yes	1.2e-005	0.063
distance WL = > 70	distance FG = -50 - 30	cover litter = < 50			Formicidae = yes	1.2e-005	0.063
distance WL = > 70	distance FG = -50 - 30	cover litter = < 60			Formicidae = yes	1.2e-005	0.063
distance FG = -50 - 30	G_1_1 = yes	height herbs = < 70			Pseudoophonus_rufipes = 1 - 2	2.4e-007	0.095
distance FG = -50 - 70	G_1_1 = yes	height herbs = < 70			Pseudoophonus_rufipes = 1 - 2	2.4e-007	0.095
G_1_1 = yes	sand = 3	height herbs = < 20			Bembidion_femoratum = 1 - 2	0.0004	0.063
distance WL = > 70	distance FG = < 90	G_2 = no	G_4 = no	cover litter = < 20	Formicidae = yes	0.00018	0.063
distance WL = > 70	distance FG = < 90	G_2 = no	G_4 = no	cover litter = < 40	Formicidae = yes	0.00018	0.063
distance WL = > 70	distance FG = < 90	G_2 = no	G_4 = no	cover litter = < 50	Formicidae = yes	0.00018	0.063
distance WL = > 70	distance FG = < 90	G_2 = no	G_4 = no	cover litter = < 60	Formicidae = yes	0.00018	0.063

Fig. 1. Example HTML output from simultaneously applying the Fisher quantifier and the lower critical implication to the terrestrial data.

Compared to contemporary data mining methods, GUHA lacks a thorough coupling to the database technology. Actually, such a coupling was under development in the 1980s [52]. However, it was oriented exclusively towards network databases relying on the Codasyl proposal [6], that time still commercially the most successful kind of databases. As Codasyl databases became obsolete, that development has been abandoned. To couple GUHA to relational and object-oriented databases remains a task for the future.

In spite of that difference, we feel that GUHA fully deserves to be considered an early example of data mining. This opinion can be justified from multiple points of view.

Purpose. In this respect, GUHA has several features typical for data mining [8,9,34,56]:

- search for relationships hidden in the data,
- limiting the search to relationships interesting according to some predefined criteria,
- focus on relationships that cannot be found in a trivial way (e.g., that could not be found through SQL queries),
- automating the search as far as possible,
- optimization to avoid blind search whenever possible.

Methods. GUHA is similar to some modern data mining approaches in employing logic for the specification of and navigation through the hypotheses space, while employing data analysis, in particular statistical methods, for the evaluation of hypotheses in that space. Moreover, that similarity goes even further, covering also the main kinds of statistical methods employed for the evaluation, namely statistical hypotheses testing, most often in the context of contingency tables [4,5,10,38,39,62,64].

Scope. GUHA relates, in particular, to *mining association rules*. Indeed, if $\mathscr{A} = \{A_1, \ldots, A_m\}$ is the set of binary attributes in a database of size k, and if $X, Y \subset \mathscr{A}$,

$X \cap Y = \emptyset$, then the association rule $X \Rightarrow Y$ is significant in the database (according to [1,35,36,40,55,61]) if and only if the GUHA sentence

$$\bigwedge_{i \in X} A_i \supset_{B,p} \bigwedge_{i \in Y} A_i$$

holds for the $k \times m$ dichotomous data matrix formed by the values of the attributes from $\mathscr{A}$. Here, $\supset_{B,p}$ is a *founded* version of the generalized quantifier $\supset_p$ mentioned in Section 2 (version requiring the frequence a to be at least as large as a predefined *base* $B \in \mathscr{N}$, see also [14,21]). Moreover, there is a very simple relationship between the parameters p and B of that quantifier, and the *support* $s \in (0,1\rangle$ and *confidence* $c \in (0,1\rangle$ of the above association rule

$$p = c \ \& \ B = k \cdot s.$$

In addition, several concepts pertaining to mining association rules have some counterpart in GUHA:

- Mining rules with *item constraints* [55] can be covered by GUHA using *relativized sentences* with filtering conditions [14,21].
- The notion of a *frontier/border set*, crucial for efficient finding of all large/frequent itemsets [44,61], is closely related to the GUHA concept of *prime sentences* [19].
- The gap between association rules and functional dependencies known from databases [1,43] can be partially bridged in GUHA by means of *improving literals* [14,19].

4. Impact of soft computing

Soft computing is a relatively new name for a branch of research including fuzzy logic, neural networks, genetic and probabilistic computing.[1] Here we contemplate on soft exploratory data analysis or soft data mining in GUHA-style and in general.

We begin with a discussion of fuzzy logic. This is admittedly a fashionable term with several meanings. Following Zadeh we shall distinguish between FL_w (fuzzy logic in wide sense) and FL_n (fuzzy logic in narrow sense), the former being practically everything dealing with fuzziness, thus synonymous with fuzzy set theory (also in wide sense). In the narrow sense, fuzzy logic is just the study of some calculi of many-valued logic understood as logic of graded truth. Zadeh stresses that the agenda of fuzzy logic differs from the agenda of traditional many-valued logic and includes entries as generalized quantifiers (usually, many, etc.), approximate reasoning and similar. In

[1] Let us quote from Zadeh, the father of fuzzy set theory and fuzzy logic [60]: "The guiding principle of soft computing is: exploit the tolerance for imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness, low solution cost and better rapport with reality. One of the principal aims of soft computing is to provide a foundation for the conception, design and application of intelligent systems employing its member methodologies symbiotically rather than in isolation".

the last period of development, fuzzy logic (FL_n) has been subjected to a serious mathematical and logical investigation resulting, among other works, in the monograph [15]. It has turned out that, on the one hand, calculi of fuzzy logic (based on the notion of a triangular norm) admit classical investigation concerning axiomatizability, completeness, question of complexity, etc., both for propositional and predicate logic, and, on the other hand several entries of Zadeh's agenda can be analyzed in terms of *deduction* in appropriate theories. The main aim of that book is to show that fuzzy logic is (can be) a real fully-fledged logic. This does not contradict the fact that fuzzy logic in wide sense has many extra-logical aspects. But mathematical foundations of fuzzy logic may be understood as an integral part of mathematical foundations of Soft Computing—and the paradigm of soft computing is obviously relevant to the intended development of DS.

5. Fuzzy logic of discovery and fuzzy GUHA

Needless to say, fuzzy logic in the wide sense has been repeatedly used in data analysis; see e.g. [2, Section 5.5], or [41] for a survey. It is very natural to ask how can the methods and results of fuzzy logic in the narrow sense be applied to the calculi of logic of discovery as sketched above; thus what are fuzzy observational and theoretical languages of DS. This should clearly not be a self-purpose fuzzification: First, the typical observational quantifiers used in GUHA are *associational* (φ, ψ are associated, positively dependent in the data) or *multitudinal* (many φ's are ψ's). Until now it has been always defined in some crisp way, using a parameter (p-many, etc.). But it is much more natural to understand them in a frame of a fuzzy logic, at least in two kinds of systems:

(a) Open observational formulas are crisp as before (like "age is $\langle 10\text{–}20 \rangle$"—yes or no), but quantified observational formulas are fuzzy, e.g. the truth value of (Many x)φ in $\mathbf{M}$ can be the relative frequence of φ in $\mathbf{M}$.

(b) Atomic observational formulas are fuzzy, as well, i.e. the attributes are fuzzy, e.g. "age is young" where "young" is a fuzzy attribute with a given fuzzy truth function on numerical values of age. For both variants, [15] contains foundations; for further development of fuzzy logic dealing with probability (and the modality "probably") see [11,16].

Second, fuzzy hypothesis testing may be developed in the framework of fuzzy generalized quantifiers in the FL_n sense, mentioned in the preceding section. Fuzziness can enter a statistical test mainly in the following ways:

(i) The data analyst has only a vague idea about the null hypothesis to test. In that case, the set $\mathscr{D}_0 \subset \mathscr{D}$, considered in Section 2, should be replaced by an appropriate fuzzy set $\tilde{\mathscr{D}}$ on $\mathscr{D}$. For example, the set $(0, \theta\rangle$ determining the null hypothesis for the parameter $p_{\psi|\varphi}$ should be replaced by a fuzzy set on $(0, 1)$.

(ii) The data analyst has only a vague idea about the critical region to use for the test. Then a fuzzy set on $(0, 1)$ should be used instead of the interval $(0, \alpha\rangle$. This corresponds to the situation when the data analyst is not sure about the significance level α to choose.

Elbe river groyne fields ecology - DB "biodat10"

fuzzy lower critical implication: truth grades for 4 definitions of "p is high", above the threshold 0.95 (according to the definition 1)

				**Species**	**Ecological factors**	
1	1	1	0.91892	Cladocera	flow velocity = 50-70 cm/s	
0.98154	0.9997	0.9943	0.65362	Copepoda	grain diameter - F hrb ter's method = 0.4-0.6 mm	
1	1	1	0.91892	Copepoda	flow velocity = 50-70 cm/s	
0.98154	0.9997	0.9943	0.65362	Nais	grain diameter - F hrb ter's method = 0.4-0.6 mm	
1	1	1	0.91892	Nais	flow velocity = 50-70 cm/s	
1	1	1	0.91892	Robackia	flow velocity = 50-70 cm/s	
0.97816	0.99858	0.97911	0.73121	Tubificidae	glowable proportion = below 10 %	
1	1	1	0.91892	Cladocera	glowable proportion = below 10 %	flow velocity = 50-70 cm/s
1	1	1	0.91892	Copepoda	glowable proportion = below 10 %	flow velocity = 50-70 cm/s

Fig. 2. Top of a HTML output from applying the fuzzy lower critical implication to the aquatic data for four particular choices of a fuzzy set capturing the notion "is high".

In our opinion, especially the fuzzification of the tested null hypotheses is highly relevant for exploratory data analysis and data mining. In fact, exploratory analysis and data mining are typically performed in situations when only very little is known about the distribution of the random variates that generated the data. Consequently, it is very difficult to specify precisely the set $\mathscr{D}_0$ determining the tested null hypothesis, e.g., to choose a precise value of the threshold θ in our example.

Statistical tests with fuzzy null hypotheses have been intensively studied in the context of the GUHA generalized quantifiers lower and upper critical implication mentioned in Section 2 [27,28]. The investigations were mainly intended for the fuzzy null hypotheses paraphrased as "$p_{\psi|\varphi}$ is not high" (replacing $p_{\psi|\varphi} \leqslant \theta$) and its respective alternative hypothesis "$p_{\psi|\varphi}$ is high" in the case of the lower critical implication, or for the null hypothesis "$p_{\psi|\varphi}$ is not low" (replacing $p_{\psi|\varphi} \geqslant \theta$) and the alternative "$p_{\psi|\varphi}$ is low" in the case of the upper critical implication. However, actually a much more general setting of nonincreasing/nondecreasing linguistic quantifiers (in the sense introduced by Yager in [59]) has been used. Recently, a first implementation of the fuzzy lower critical implication has been finished [31]. An example of results obtained with this fuzzy generalized quantifier in the ecological application introduced in Section 3 is given in Fig. 2.

A number of important results concerning the fuzzy-hypotheses generalizations $\supset^{!}_{\sim}$, $\supset^{?}_{\sim}$ of the quantifiers $\supset^{!}$, $\supset^{?}$, respectively, have been proven in [28]. Logical theory of the fuzzy quantifiers "the probability of … is high" and "the conditional probability of … given … is high" is elaborated in [15].

All this seems to be a promising research domain. Let us add two details:

Languages for data concerning event sequences should be developed, i.e. the ordering of objects in the data matrix is relevant and expressible in the language. Some rudimentary beginnings can be found in [19].

For processing extremely large data sets by a GUHA-like procedure, there are good possibilities of parallelization.

6. Conclusion

This paper is a position paper. We have offered formal logical foundations (partly old and forgotten, partly new and under development) for a certain direction of discovery science, namely logic of discovery as hypothesis formation. This is based on fully fledged formal calculi with exactly defined syntax and semantics in the spirit of modern mathematical logic. We have stressed the paradigm of soft computing, in particular of fuzzy logic together with its strictly logical foundations. In our opinion, the surveyed kind of logic of discovery is highly relevant to the objectives of DS.

Acknowledgements

This paper is a revised version of our contribution presented at the First International Conference on Discovery Science, 1998 [20]. Its revision was supported by the grant IAA 1030004 of the Grant Agency of the Academy of Sciences of the Czech Republic.

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proc. 20th Internat. Conf. on Very Large Data Bases, 1994.
[2] C. von Altrock, Fuzzy Logic and Neurofuzzy Applications Explained, Prentice-Hall, PTR, Upper Saddle River, NJ, 1995.
[3] B.G. Buchanan, Logics of Scientific Discovery, Stanford AI Memo No. 47, Stanford University 1966.
[4] C. Chatfield, Model uncertainty, data mining and statistical inference, J. Roy. Statist. Soc. Ser. A 158 (1995) 419–466.
[5] T. Chau, A.K.C. Wong, Pattern discovery by residual analysis and recursive partitioning, IEEE Trans. Knowledge Data Eng. 11 (1999) 833–852.
[6] CODASYL Data Base Task Group, DBTG Report, Technical Report, ACM, 1971.
[7] J. Doubravová, A. Sochorová, Testing interpersonal hypothesis of music using GUHA method. Languages Des., 1996.
[8] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery: an overview, in: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 1–36.
[9] W. Frawley, G. Piatetsky-Shapiro, C. Matheus, Knowledge discovery in databases: an overview, in: G. Piatetsky-Shapiro, W. Frawley (Eds.), Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA, 1991, pp. 1–27.
[10] C. Glymour, D. Madigan, D. Pregibon, P. Smyth, Statistical inference and data mining, Comm. ACM 39 (1996) 35–41.
[11] L. Godo, F. Esteva, P. Hájek, Reasoning about probability using fuzzy logic, Neural Network World 10 (2000) 811–824.

[12] P. Hájek, On logics of discovery, in: Mathematical Foundations of Computer Science, Lecture Notes in Computer Science, Vol. 32, 1975, Springer, Berlin, pp. 30–45.
[13] P. Hájek, Decision problems of some statistically motivated monadic modal calculi, Internat. J. Man-Mach. Stud. 15 (1981) 351–358.
[14] P. Hájek, The new version of the GUHA procedure ASSOC (generating hypotheses on associations)—mathematical foundations, in: COMPSTAT 1984—Proc. in Computational Statistics, 1984, pp. 360–365.
[15] P. Hájek, Metamathematics of Fuzzy Logic, Kluwer Academic Publishers, Dordrecht, 1998.
[16] P. Hájek, F. Esteva, L. Godo, Fuzzy logic and probability, in: P. Bernard, S. Hanks (Eds.), Uncertainty and Artificial Intelligence '95, Morgan Kaufmann, San Mateo, 1995, pp. 237–244.
[17] P. Hájek, I. Havel, M. Chytil, The GUHA-method of automatic hypotheses determination, Computing 1 (1966) 293–308.
[18] P. Hájek, T. Havránek, On generation of inductive hypotheses, Internat. J. Man-Mach. Stud. 9 (1977) 415–438.
[19] P. Hájek, T. Havránek, Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer, Berlin, Heidelberg, New York, 1978; also www.cs.cas.cz/∼hajek/guhabook.
[20] P. Hájek, M. Holeňa, Formal logics of discovery and hypothesis formation by machine, in: S. Arikawa, H. Motoda (Eds.), Discovery Science, Springer, Berlin, Tokyo, 1998, pp. 291–302.
[21] P. Hájek, A. Sochorová, J. Zvárová, GUHA for personal computers, Comput. Statist. Data Anal. 19 (1995) 149–153.
[22] J. Hálová, O. Štrouf, P. Žák, A. Sochorová, N. Uchida, T. Yuzuvi, K. Sakakibava, M. Hirota, QSAR of catechol analogs against malignant melanoma using fingerprint descriptors, Quant. Struct.-Act. Relat. 17 (1998) 37–39.
[23] D. Harmancová, M. Holeňa, A. Sochorová, Overview of the GUHA method for automating knowledge discovery in statistical data sets, in: M. Noirhomme-Fraiture (Ed.), Knowledge Extraction and Symbolic Data Analysis, Eurostat, Luxembourg, 1999, pp. 65–77.
[24] T. Havránek, The approximation problem in computational statistics, in: J. Bečvář (Ed.), Mathematical Foundations of Computer Science '75, Lecture Notes in Computer Science, Vol. 32, 1975, pp. 258–265.
[25] T. Havránek, Statistical quantifiers in observational calculi: an application in GUHA method, Theory and Decision 6 (1975) 213–230.
[26] T. Havránek, Towards a model theory of statistical theories, Synthese 36 (1977) 441–458.
[27] M. Holeňa, Exploratory data processing using a fuzzy generalization of the GUHA approach, in: J. Baldwin (Ed.), Fuzzy Logic, Wiley, New York, 1996, pp. 213–229.
[28] M. Holeňa, Fuzzy hypotheses for GUHA implications, Fuzzy Sets and Systems 98 (1998) 101–125.
[29] M. Holeňa, Traditional and modern artificial intelligence explores ecological data, in: H. Hyötyniemi, (Ed.), STeP 2000: Millenium of Artificial Intelligence, 2000.
[30] M. Holeňa, Observational logic integrates data mining based on statistics and neural networks, in: D.A. Zighed, J. Komorowski, J.M. Żytkov (Eds.), Principles of Data Mining and Knowledge Discovery, Springer, Berlin, 2000, pp. 440–445.
[31] M. Holeňa, A fuzzy logic framework for testing vague hypotheses with empirical data, in: Proc. Fourth Internat. ICSC Symp. on Soft Computing and Intelligent Systems for Industry, ICSC Academic Press, Sliedrecht, 2001 pp. 401–407.
[32] M. Holeňa, Statistical, logic-based, and neural networks based methods for mining rules from data, in: A.K. Hyder, V. Bystritskii (Eds.), Multisensor and Sensor Data Fusion, NATO Science Series Publishers, in preparation.
[33] M. Holeňa, A. Sochorová, J. Zvárová, Increasing the diversity of medical data mining through distributed object technology, in: P. Kokol, B. Zupan, J. Stare, M. Premik, R. Engelbrecht (Eds.), Medical Informatics Europe '99, IOS Press, Amsterdam, 1999, pp. 442–447.
[34] M. Holsheimer, A. Siebes, Data mining, The search for knowledge in databases, Technical Report, CWI, Amsterdam, 1994.
[35] M. Houtsma, A. Swami, Set-oriented mining of association rules, Technical Report, IBM Almaden Research Center, 1993.
[36] M. Kamber, J. Han, J. Chiang, Using data cubes for metarule-guided mining of multi-dimensional association rules, Technical Report, Database Systems Research Laboratory, Simon Fraser University, 1997.

[37] U.J. Keisler, Probability quantifiers, in: J. Barwise, S. Feferman (Eds.), Model-Theoretic Logics, Springer, New York, 1985, pp. 539–556.
[38] W. Klösgen, Efficient discovery of interesting statements in databases, J. Intell. Inform. Systems 4 (1995) 53–69.
[39] W. Klösgen, Explora: a multipattern and multistrategy discovery assistant, in: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 249–272.
[40] F. Korn, A. Labrinidis, Y. Kotidis, C. Faloutsos, Quantifiable data mining using ration rules, VLDB J. 8 (2000) 254–266.
[41] R. Kruse, K.D. Meyer, Statistics with Vague Data, Reidel, Dordrecht, 1987.
[42] D.I. Lin, Z. Kedem, Pincer search: a new algorithm for discovering the maximum frequent set, in: Proc. EDBT'98: Sixth Internat. Conf. on Extending Database Technology, 1998.
[43] H. Mannila, K. Räihä, Dependency inference, in: Proc. 13th Internat. Conf. on Very Large Data Bases, 1987, pp. 155–158.
[44] H. Mannila, H. Toivonen, I. Verkamo, Efficient algorithms for discovering association rules, in: U. Fayyad, R. Uthurusamy (Eds.), Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA, 1994, pp. 181–192.
[45] A. Mueller, Fast sequential and parallel algorithms for association rule mining: a comparison, Technical Report, Department of Computer Science, University of Maryland, College Park, 1995.
[46] L. Pecen, K. Eben, Non-linear mathematical interpretation of the oncological data, Neural Network World 6 (1996) 683–690.
[47] L. Pecen, E. Pelikán, H. Beran, D. Pivka, Short-term fx market analysis and prediction, in: Neural Networks in Financial Engeneering, 1996, pp. 189–196.
[48] L. Pecen, N. Ramešová, E. Pelikán, H. Beran, Application of the GUHA method on financial data, Neural Network World 5 (1995) 565–571.
[49] G.D. Plotkin, A further note on inductive generalization, Mach. Intell. 6 (1971) 101–124.
[50] K.R. Popper, The Logic of Scientific Discovery, Hutchinson, London, 1974.
[51] P. Pudlák, F. Springsteel, Complexity in mechanizing hypothesis formation, Theoret. Comput. Sci. 8 (1979) 203–225.
[52] J. Rauch, Logical problems of statistical data analysis in data bases, in: Proc. 11th Seminar on Data Base Management Systems, 1988, pp. 53–63.
[53] J. Rauch, Logical calculi for knowledge discovery in databases, in: Komarowski, Żytkov (Ed.), Principles of Data Mining and Knowledge Discovery, Lecture Notes in AL, Vol. 1263, Springer, Berlin, 1997.
[54] J. Rauch, Classes of four-fold table quantifiers, in: Quafafou, Żytkov (Eds.), Principles of Data Mining and Knowledge Discovery, Springer, Berlin, 1998, pp. 203–210.
[55] R. Srikant, Q. Vu, R. Agrawal, Mining association rules with item constraints, in: Proc. Third Internat. Conf. on Knowledge Discovery and Data Mining, KDD-97, 1997.
[56] A. Teller, M. Veloso, Program evolution for data mining, Internat. J. Expert Systems 8 (1995) 216–236.
[57] H. Toivonen, Discovery of frequent patterns in large data collections, Ph.D. Thesis, University of Helsinki, 1996.
[58] J.W. Tukey, Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.
[59] R. Yager, On a semantics for neural networks based on fuzzy quantifiers, Internat. J. Intell. Systems 7 (1992) 765–786.
[60] L.A. Zadeh, What is soft computing? (Editorial) Soft Comput. 1 (1997) 1.
[61] M. Zaki, S. Parathasarathy, M. Ogihara, W. Li, New parallel algorithms for fast discovery of association rules, Data Mining Knowledge Discovery 1 (1997) 343–373.
[62] R. Zembowicz, J. Żytkov, From contingency tables to various forms of knowledge in databases, in: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 329–352.
[63] J. Zvárová, J. Preiss, A. Sochorová, Analysis of data about epileptic patients using Guha method, Internat. J. Medical Inform. 45 (1997) 59–64.
[64] J. Żytkov, R. Zembowicz, Contingency tables as the foundation for concepts, concept hierarchies and rules: the 49er system approach, Fund. Inform. 30 (1997) 383–399.